Synthese (2024) 203:90
https://doi.org/10.1007/s11229-024-04508-z

ORIGINAL RESEARCH

Reference in remembering: towards a simulationist account

James Openshaw[1] · Kourken Michaelian[1,2]

Received: 18 June 2023 / Accepted: 18 January 2024

Abstract

Recent theories of remembering and of reference (or singular thought) have de-emphasised the role causation was thought to play in mid- to late-twentieth century theorising. According to postcausal theories of remembering, such as simulationism, instances of the psychofunctional kind *remembering* are not, in principle, dependent on appropriate causal chains running from some event(s) remembered to the occurrence of remembering. Instead they depend only on the reliability, or proper functioning, of the cognitive system responsible for their production. According to broadly reliabilist accounts of *singular thought*, such thought is not, in principle, dependent on causal chains running from the object(s) of thought to the occurrence of thinking. Despite this common trend, accounts of the two phenomena have been pursued separately. In this paper, we argue that the two lines of research can profitably converge to address a neglected question: what enables occurrences of remembering to *refer* to particular events, and what determines *which* event a given occurrence of remembering refers to? We motivate and present a reliabilist account of reference-fixing for postcausal theories of remembering, focusing in particular on simulationism. We then show that this account draws attention to the possibility of *referential mnemic confabulation:* cases where the reliability requirement for reference is met despite the improper functioning of the episodic construction system. We suggest that this makes sense of some underdiscussed phenomena described in the empirical literature on confabulation and argue that our reliabilist account of mnemic reference-fixing accommodates these more naturally than could causal theories of remembering.

Keywords Episodic memory · Reference · Remembering · Simulationism · Confabulation

✉ James Openshaw
jamesopenshaw0@gmail.com

[1] Centre for Philosophy of Memory, IPhiG, Université Grenoble Alpes, Grenoble, France

[2] Institut Universitaire de France, Paris, France

Published online: 05 March 2024

1 Introduction

Two landmark contributions to the study of memory and of reference—Martin and Deutscher's 'Remembering' (1966) and Kripke's 1970 'Naming and Necessity' lectures—generated a seismic shift towards causal accounts of remembering and singular thought (and away from epistemic and descriptive accounts, respectively). In recent years, theories of the two phenomena have placed less emphasis on the role of causation. Simulationists (Michaelian, 2016b) claim that instances of the psychofunctional kind *remembering* are, in principle, not dependent on appropriate causal chains running from some event(s) remembered to the occurrence of remembering. Instead they depend only on the reliability, or proper functioning, of the cognitive system responsible for their production. Analogously, broadly reliabilist accounts of *singular thought* suggest that instances of such thought are, in principle, not dependent on causal chains running from the object(s) of thought to the occurrence (or state) of thinking (Dickie, 2015).[1] In the case of memory, the rationale is a growing, empirically-driven awareness that the content of a retrieved representation will seldom be exclusively related by a content-transmitting causal chain to a particular past experience, and distributed conceptions of memory traces only heighten this awareness. In the case of singular thought, the rationale is a growing concern that broadly causal accounts of perception- and testimony-based singular thoughts cannot be applied to all cases (Hawthorne & Manley, 2012; Jeshion, 2010). Despite this common trend, accounts of remembering and of singular thought have been pursued separately.

In this paper we argue that the two lines of research can profitably converge to make progress on a question which has seldom been given focused attention: under what conditions is there some event in one's personal past *e* such that one is remembering *e?* In other words: what enables occurrences of remembering to *refer* to particular events, and what determines *which* event a given occurrence of remembering refers to?

While causal theories of remembering (Bernecker, 2010; Martin & Deutscher, 1966; Werning, 2020) may seem to have an easy route to answering this question, we summarise some grounds for concern (§2). We then clarify the existing resources available to recent 'postcausal' theories of remembering, focusing in particular on simulationism (§3). After finding these wanting (§4), we articulate a broadly reliabilist theory of mnemic reference-fixing (§54). The principal upshot of our discussion is that postcausal theories have the beginnings of a plausible solution to the challenge about how reference gets fixed. A secondary upshot is that this answer draws attention to the possibility of cases of *referential mnemic confabulation:* cases where the reliability requirement for reference is met despite the improper functioning of the subject's episodic construction system. We suggest that this makes sense of some underdiscussed phenomena described in the empirical literature on mnemic confabulation (§6)

[1] Dickie (2015) is less concerned with capturing cases that intuitively involve reference but don't involve the right sort of causal chains, and more with the explanatory question as to *why* causal chains allow us to achieve reference. The answer is that it is because they allow us to achieve 'cognitive focus'. Nevertheless, it is arguably one of the virtues of the kind of view Dickie (2015) articulates that, by getting the right generality of explanation, we can in principle extend the theory beyond causal cases.

and that our reliabilist theory of mnemic reference-fixing accommodates these more naturally than could causal theories.

2 Reference and episodic memory

Episodic memory (Tulving 1972) enables us to consciously ‘relive’ experienced events from our personal past. For example, you may remember making coffee this morning and sensorily recollect what it was like to smell the coffee grounds or to see the kettle reach a boil. Successful remembering seems to require a certain relationship between one’s present recollection and some past event. Of course, the recollection must be suitably accurate. Yet, before questions of (in)accuracy can even arise, something must first ‘fix’ or determine that the recollection is *about that* particular occasion you brewed coffee, rather than, say, the previous morning. By analogy, success in uttering ‘This is blue’ requires, for its very evaluability, that ‘This’ refers to a particular object and, for its truth, that the predicate accurately characterises the referent. Though these observations are simple, what we might call the *reference-fixing* and *accuracy* conditions of remembering remain obscure.

In this paper, we focus on the issue of reference-fixing.[2] Compare the following two questions we can ask about remembering:

(Q1) Under what conditions does remembering (i.e., the psychofunctional kind) occur?

(Q2) Under what conditions is there some event in one’s personal past e such that one is remembering e?

The intended reading of (Q1) concerns the conditions under which the ‘narrow’ psychofunctional process type *remembering* occurs or is tokened. In contrast, (Q2) concerns the conditions under which ‘wide remembering’, or the semantic type *referential remembering*, is tokened. On some views, the conditions for referential remembering, (Q2), may be more demanding than the conditions for remembering per se, (Q1). That is, remembering may not require, in every case, that there is some event in one’s past such that one is remembering it.[3]

Whether or not one takes referentiality to be a contingent feature of remembering, the following question presents itself to all who acknowledge the mere possibility of reference in remembering (i.e., to those who do not answer (Q2) by denying that such conditions are ever met), and it is with this question that this paper is centrally concerned.

The reference question: Given that S is remembering some event(s), what determines *which* event(s) S is remembering?

[2] Of course, the two issues are closely related. The issue of accuracy has begun to receive more attention, yet there is arguably a sense in which the question of reference is more fundamental. If it is constitutive of remembering that when successful it refers to experiences in the subject’s past (Hoerl 2018), or whether instead it sometimes refers to mind-independent events or objects (Openshaw 2022), it is this issue that will determine whether *authenticism* is generally true. We discuss accuracy further in §5.

[3] Analogously, some deny that a theory of remembering must be a theory of *successful* remembering, rejecting a factive conception of the explanandum (De Brigard (2014); Michaelian (2016b, pp. 69–70)).

2.1 Causalism and reference in episodic memory

Given the enduring dominance of broadly causal theories of how the referents of proper names (Evans, 1973; Kripke, 1980) and singular thoughts (Devitt, 1981; Recanati, 2012) are determined, it would not be surprising to find a similar dominance of causal theories of mnemic reference-fixing. To the extent that one thinks of reference as always, or perhaps just paradigmatically, fixed by virtue of the presence of a causal relation between the tokening of a singular term or thought-vehicle and the referent, it is natural to be attracted to causal solutions to the reference question. As we will see, this expectation is also compounded by the influence of causal theories of remembering itself.

For instance, Recanati (2007) writes:

> Episodic memories are mental states which presuppose other mental states […] to which they are related both causally (the memory derives from the perceptual experience, which leaves it as a 'trace') and semantically (the memory inherits the content of the perceptual experience) (2007, p. 136).

Soteriou likewise suggests that an answer to the reference question need only appeal to the causal aetiology of a retrieved memory trace: "*which particular* past event is represented […] is determined by the causal ancestry of the memory" (2018, p. 308). Finally, although Werning and Liefke (forthcoming) deny that memory traces—or what they call 'minimal traces'—carry representational content, they do claim: "the primary experience that underlies a particular episodic memory must be uniquely identifiable. This can be achieved by a minimal trace alone" (p. 32). A recurring suggestion, then, is that the reference question can be answered, indeed *only* answered, by appealing to the causal-informational relation embodied in the theoretical notion of a memory trace. Although the majority of theorists are now at pains to emphasise that the content of remembering is "not a literal reproduction of the past" but the result of "a constructive process in which bits and pieces of information *from various sources* are pulled together" (Schacter & Addis, 2007, p. 773; emphasis added), the idea that *which* event one remembers (i.e., the referent) is determined by the causal source of a *privileged* ingredient continues to hold influence.

At least since Martin and Deutscher's (1966) classic paper, it has commonly been thought that an analysis of what remembering itself is should fundamentally involve the specification of an appropriate causal relation. Their proposal is summarised as follows:

> If someone *remembers something*, whether it be 'public,' such as a car accident, or 'private,' such as an itch, then the following criteria must be fulfilled:
>
> 1. Within certain limits of accuracy he represents *that past thing*.
> 2. If the thing was 'public,' then he observed what he now represents. If the thing was 'private,' then it was his.
> 3. His past experience of the thing was operative in producing a state or successive states in him finally operative in producing his representation (Martin & Deutscher, 1966, p. 166; emphasis added).

Notice that here, and throughout their paper, what Martin & Deutscher are answering is in fact either simply (Q2), or, more plausibly, (Q1) and (Q2) simultaneously. If the latter, then the background assumption is that remembering is a 'wide' psychofunctional state, individuated in part by its referentiality: unless you are referring to something in your episode of remembering, you are not remembering at all. As Hopkins (2018) puts in, in the form of a rhetorical question, after stating that "[e]pisodic memories always have singular content": "how can you be having an episodic memory unless there is an answer to the question which episode you are remembering?" (2018, p. 60). We think that something like this traditional *causal theory of remembering* (CTM) has often seemed inevitable precisely because it has been presupposed that a theory of *remembering* must be a theory of 'wide remembering', and so provide an answer to (Q2) and therefore to the reference question.[4]

Omitting the idiosyncratic details of specific theories, the causalist claims:

(CTM): S remembers event *e* if and only if

(i) S now represents *e*
(ii) S experienced *e* when it occurred
(iii) there is an *appropriate* causal connection, i.e. one sustained by a memory trace, between the subject's original experience of *e* and her retrieved representation of *e*.[5]

A key part of this theory is that memory traces not only carry some form of content, they are *discriminating* in that they 'point back' to the particular events on exposure to which they were originally formed. To use Langland-Hassan's (2022) terminology (although he in fact defends a non-traditional form of causalism), memory traces are 'monogamous' by nature. If they are often overwritten with new content (e.g., in reconsolidation), it is not in a way that typically destroys the uniqueness of their origin-determined reference. What matters for causalists is the presence of a one–one (or perhaps many-one) appropriate causal relation between memory trace and experienced event. And appropriate causation can then play a two-part role: (i) distinguishing remembering from relearning (see footnote 5); (ii) serving as a mnemic reference-fixer. If *S* is remembering, *what S* remembers is simply the event that gave rise to the experience that then (uniquely) gave rise to the memory trace now 'finally operative in producing' S's current representation. If part of what it is to be remembering is for there to be an appropriate causal relation to an event in one's past, then *what* one counts as remembering *is that past event*. For causalists, then, (Q1) and (Q2) are practically inseparable, and an answer to the reference question simply falls out of their account of the conditions under which remembering occurs.

[4] We consider some of the implications of holding that remembering is constitutively referential below.

[5] As Martin & Deutscher put this final clause, the subject's original experience of *e* must be "operative in producing the state (or successive set of states) in him which is finally operative in producing the representation *in* the circumstances in which he is prompted" (1966, p. 185). The salient point is that, to avoid collapsing the intuitive distinction between remembering and *relearning* (1966, 180ff), the memory trace is both produced by the subject's experience of *e* and is operative in his recollection.

2.2 Postcausalism and reference in episodic memory

Although causalist answers to the reference question are not our focus, we wish to register some concerns about the apparent ease with which causalists can provide an answer to the reference question. We don't think that there is *any* theory of remembering which at present has a clearly satisfying answer to the reference question. First and foremost, there are question marks hanging over the naturalistic credentials of any view of remembering which requires, as a matter of necessity, that a subject is only remembering an event if there is an appropriate causal connection of the sort described above (see, e.g., Andonovski (2022)). Relatedly, whether there is an episodic *memory* system with a proprietary store of information eligible to play the memory trace role in general is an open question in psychology (Addis, 2020; Rubin, 2022). Second, it could turn out that the kind of causation that is'appropriate' for remembering is weaker than the kind of causation that is 'appropriate' for successful reference. In that case, even causalist views would have to say more about the distinctive variety of reference involved in remembering. One way this might manifest is if the causal aetiology of memory traces is often insufficiently discriminating (Langland-Hassan (2022); Michaelian (2021, 7483ff); Robins (2016)), so that simply retrieving a memory trace does not guarantee that one is appropriately causally related to a single event rather than to multiple events[6]. Finally, problems facing causal theories of reference in general will be inherited by simple causal answers to the reference question. For instance, it is not easy to identify causal-historical relations that enable one to distinguish between reference to statues versus lumps of clay, temporally extended 'worms' versus momentary 'stages', or whole objects versus 'undetached parts' (for relevant discussion see Deutsch (2021), Sterelny (1990), and Williams (2008)). Prima facie, then, it will not be easy to identify causal-historical relations that enable one to distinguish between reference to *events* involving one over any other such entities.

In spite of these concerns, there may seem to be special challenges raised by the reference question for recent theories that reject the necessity of appropriate causation. According to *postcausal* theories, remembering an event does not necessarily require an appropriate causal connection to it (however the details of 'appropriateness' are ultimately be cashed out). If appropriate causation cannot perform the task of acting as a mnemic reference-fixer in all cases of remembering, then what can? The postcausal theory on which we focus is Michaelian's (2016b) simulationism. And so the question in what follows becomes: can simulationists provide an adequate solution to the reference question?

Before outlining simulationism (§3), rejecting some 'easy' answers (§4), and arguing for a new account of reference in remembering (§5), we wish to note the pressing nature of the reference question for other recent theories and to highlight the salience of our discussion beyond simulationism. Given our proposed way of carving up the landscape, which views fall under 'postcausalism' will vary with one's interpretation of the term 'appropriate causation'.[7] One view that deserves mention is Fernández's

[6] For an in-depth account of this problem for 'pure' causal theories of mnemic reference, and a parallel discussion of alternatives, see Barkasi (forthcoming), in this topical collection.

[7] By contrast, Michaelian and Robins (2018: 23) define postcausalism as the view that no causal connection, *whether 'appropriate' or not*, is necessary for remembering. So our usage is more liberal.

(2019) functionalist theory of remembering. According to his view, what makes a token mental state (or occurrence) an instance of remembering is that it plays the right functional role, this role being that it *tend* to be appropriately caused. Fernández will therefore presumably need a theory of reference in cases where things do not go as they tend to go, i.e., when there is no appropriate causation. Another view which is perhaps not easy to classify here is Perrin's (2021) procedural causalism. It requires, in place of (iii) in (CTM), that S's current representation of *e* is caused by a procedural pattern originating in S's original experience of *e*, where a procedural pattern is a form of "motor information", transmitted from the time of S's original experience of *e*, which is "not included into the imagistic content [of the memory] but on which S draws to reconstruct that content" (p. 240). As Perrin recognises, re-enacting past experience on the basis of a procedural pattern might often be insufficiently discriminating to secure reference to particular events (2021, p. 247, n. 27). So this theory, too, will need to be supplemented. The account we go on to sketch in §5 may provide useful resources for these and other theories as well as for simulationism.

3 Simulationism and reference in episodic memory

The simulation theory of remembering (STM) (Michaelian, 2016b) is chiefly motivated by empirical evidence that appears to cast doubt on (CTM). Briefly, research on the *constructive* character of remembering suggests that there is neither a one–one relationship between memory traces and experienced events nor between memory traces and genuine memory representations. Multiple experiences will typically leave their mark on any single trace, and there is no naturalistically motivated reason to expect that all genuine memory representations must have content transmitted via a memory trace rather than via less causally discriminating sources. Moreover, *mental time travel* research suggests that the same broadly imaginative process, carried out by the same constructive, neurocognitive system, is at work both when we imagine the future and when we remember the past. This process trivially does not involve content-transmission in the future-oriented case, and this suggests that it need not involve content-transmission in the past-oriented case, either. The simulationist takes these trajectories to their natural conclusion, proposing that remembering is distinguished from future-directed and counterfactual simulation only in being a *past-oriented* output of the same episodic construction system (ECS).[8]

(STM): S remembers event *e* if and only if

(i) *S* now represents *e*
(ii) S's current representation of *e* is produced by a properly functioning and hence reliable ECS that aims to produce a representation of an event belonging to S's personal past (Michaelian, forthcoming, p. 2).

Since it will be important later, it is worth emphasising that the notion of reliability as it appears in (ii) does not concern the frequency with which the system outputs

[8] The so-called *continuist–discontinuist* debate (Perrin 2016) is in the background in this paper, since it does not perfectly overlap the causalist–postcausalist debate which frames our discussion.

generally accurate representations. Rather, it concerns the tendency of the specific retrieval process used on that occasion to produce generally accurate representations. An individual who routinely confabulated might nonetheless, on some occasion(s), genuinely remember (Michaelian, 2020, p. 146). What matters is that the specific process by which one's purported memory representation was produced *on that occasion* tends to result in a sufficiently high degree of accuracy. In the first instance, then, reliability is a property of process *types* rather than process tokens. Yet a process token can count as reliable in a derivative sense if the type of which it is a token is reliable. The processes in question will need to be fairly fine-grained to play the relevant role in determining whether a subject is remembering. But there need not be multiple *actual* instances of a type in order for it to be reliable or unreliable. Reliability is in this sense a modal notion.

At face value, the material on the left-hand side of the biconditional in (STM) suggests that the simulationist is offering an analysis of the same phenomenon as (CTM). But on closer inspection things are less clear. It is unclear whether what is being analysed is simply what it takes for there to be an occurrence of the psychofunctional process *remembering* or instead what it takes for there to be some event such that one is remembering *it*. In other words, it's unclear whether this is an analysis that answers the 'narrow' question (Q1), the 'wide' question (Q2), or both. To illustrate, here are two distinct answers that (STM) may be seen as providing, to (Q1) and (Q2), respectively.

(STM-1): One is remembering if and only if one has a representation produced by a properly functioning and hence reliable ECS that aims to produce a representation of an event from one's personal past.

(STM-2): There is some event in one's personal past *e* such that one is remembering *e* if and only if one has a representation *referring to e* produced by a properly functioning and hence reliable ECS that aims to produce a representation of an event from one's personal past.

If the core proposal of simulationism as a theory of remembering is to provide a theory of what remembering *is*, with the semantic notion of reference—like the normative notions of accuracy or success (Michaelian, 2016b, pp. 69–70)—being a contingent feature, then they should be understood as answering only (Q1) and making only the claim in (STM-1). What they should say in response to (Q2) is a question left open by their theory. Of course, the simulationist like almost anyone will wish to answer (Q2) at some point. But, as with the conditions for accuracy, devising a theory of the conditions for reference in remembering will be conceived of as a largely separable task.

If the simulationist is answering (Q2) after all, and asserting (STM-2), their answer raises an obvious question: *What does it take to have a representation so-produced that is of some particular event?* The reference question must be addressed.

In the coming sections, we will explore possible elaborations of (STM-2) that are available to the simulationist. We will set aside the question of whether the simulationist should take reference to be a contingent or a constitutive feature of remembering; that is, whether the elaborations will be part of the simulationists answer to (Q2) only or whether the elaboration will be part of the simulationist's basic claim about the conditions for remembering per se. While we think either approach could be taken up, the former is more congruous with the simulationist's pointedly descriptive, naturalistic project (McCarroll et al., 2022), as Openshaw (2023) argues. Either way, clearly separating (Q1) and (Q2), and clearly demarcating (STM-1) and (STM-2), promises to bring clarity to the recent literature on remembering. These two questions must not be confused for one another, and disputes between parties who are addressing different questions should be avoided.

Before exploring the possible elaborations of (STM-2), it is worth noting that simulationists cannot simply retreat from answering (Q2), deny reference altogether, and embrace a kind of *anti-referentialist* view. If the ECS can be assessed for reliability, as (STM) claims, its outputs must sometimes be assessable for accuracy.[9] And they can only be assessed for accuracy if they refer. Roughly, reliability is a matter of a process's tendency to produce accurate representations. And in the absence of a subject matter (reference), a representation cannot be evaluated for accuracy. So even if simulationism is in principle compatible with the claim that *memory traces* are contentless (Michaelian & Sant'Anna 2021), it is not obviously compatible with the radical enactivist's more general hostility to mental content (Hutto & Myin, 2013) or with a local hostility to the contentfulness of episodic memory. In other words, even if the simulationist is free to deny the referential character of the vehicles involved in encoding, storage, or retrieval throughout the episodic memory process, they are not free to deny the referential character of the ultimate *output* of the episodic memory process. At the end of the process, there must (sometimes) be accuracy-evaluable content, and—if we are right about the referential character of successful remembering—that means there must be referential content.

4 Reference for simulationists: some unpromising accounts

This section briefly considers several existing approaches or quick solutions. It argues that none of them can offer a satisfying account of reference for simulationists. This motivates the pursuit of a new approach, set out in §5.

4.1 Referential heterogeneity/pluralism

Supposing that the simulationist faces up to the reference question and asserts (STM-2), it might nonetheless be that they should ultimately decline to give an informative, *general* account of what it is to have a representation, referring to something, that was produced in the relevant way. That is, they might deny that an account of what it is for a subject to be remembering commits one to providing a uniform explanation as to how it is determined which thing it is that the subject is remembering. The right-hand side of (STM-2) commits the simulationist to providing some account of what it is to have a representation referring to something that was produced in the appropriate

[9] There are other stories one could tell about the notion of *proper function* or *reliability* that may not require that the system produce representations evaluable for accuracy. For example, perhaps certain teleological theories could be adopted. Given our constraints in this paper, we set these aside and assume a broadly *process reliabilist* framework.

way by one's ECS. But suppose that there is nothing uniform to be said about what it takes to have a representation produced in the simulationist's special way that is about a particular thing. Perhaps some cases involve 'appropriate causation' while many others involve something much more circuitous and opportunistic. In that case, it might be that the most general thing we can say about the nature of remembering is simply (STM-2). Since remembering turns out to place no special constraints on the representation of the thing remembered, the simulationist need not give an account of reference-fixing *for remembering*. There is no distinctive or interesting answer to the reference question. Call this the *heterogeneity view*.

We think this view is much too pessimistic at this stage of inquiry. As a last resort, the heterogeneity view might be tolerable. But it should not be anyone's opening gambit.

A less pessimistic way of taking a broadly pluralist approach to reference-fixing, however, would be to suggest that there *is* a distinctive story to be told about how the right-hand side of (STM-2) is to be cashed out, but that this story bottoms out in a compact set of genuinely explanatory but nonetheless distinct stories. According to the *pluralist view* of mnemic reference-fixing, what explains how reference to an event is determined varies according with features of the subject's situation. For example, appropriate causation via a memory trace might fix reference in a large number of cases, whereas in others a different story must be told (see below). Perhaps some cases will involve the presence of multiple reference-fixers, each of which plays some role, with the ultimate referent being the product of some weighted average. Although less pessimistic in spirit, we think this approach has its costs, too. For one thing, there is no guarantee that the different reference-fixers will not sometimes pull in very different directions, so that 'averaging' cannot deliver the right results or, worse, one comes out as simultaneously remembering multiple very different events. Dialectically speaking, the story also runs the risk of giving simulationists too little to say about reference in remembering, encouraging a suspicion that they may not have a systematic story to tell. Pluralism, too, then, has its costs.[10]

4.2 Cue-based reference-fixing

In this sub-section we briefly review and evaluate Michaelian's (2016b) existing suggestion about how reference in remembering is achieved. Condition (ii) of (STM) contains the requirement that one's current representation is a product of an ECS which 'aims to produce a representation of an event belonging to S's personal past'. So what determines which target is being aimed at? Upon considering the reference question, he suggests:

> The obvious mechanism to which we might appeal is intention—either the intention of the subject himself or the 'intention' of his episodic construction system. In light of the possibility that the subject might misclassify another form of episodic imagination as memory [i.e., since it should be possible for a subject to intend and to think that she is remembering something when she isn't], the latter

[10] Thanks to André Sant'Anna and Denis Perrin for comments on this approach. We revisit the pluralist approach briefly in §5.2.

> option is preferable. Again, while it might seem odd to think of the episodic construction system as intending to simulate a determinate episode from the past, *this can be understood as shorthand for talk of the system responding to given retrieval cues* provided by either the agent or his environment (e.g., 'how did I get home yesterday?') (2016b, p. 112; emphasis added).

The key feature of this view is that reference is determined *before* the episodic construction process begins and, in particular, that it is determined by the presence of retrieval cues bearing some relation (perhaps not always a relation of the same kind) to the past event(s) thereby remembered. We call this the *cue-based* approach to reference.

It is natural to worry about how this sort of view could handle cases of involuntary or 'unbidden' remembering (particularly if Berntsen (2021, p. 2) is right about their frequency) where it is not obvious that there is anything both (i) 'cue'-like and (ii) referential (Michaelian, 2021, p. 7498). Without any assurance that in every case in which a subject remembers some event(s) *e* there will be some reference-giving prompt, it is doubtful that this sort of view can give us a complete or fully general answer to the reference question. One response would be to embrace the claim that cases of involuntary remembering lack referents. But there is no independent reason to think that this should be the case. Moreover, it is natural to think that one might initially be driven by retrieval cue *c*, bearing a unique relation to event *e*, to remember and yet, for one reason or another, accidentally wind up remembering some distinct event *e'*. But on this approach there is no space for mistakes of this sort to occur. The task to be taken up in the remainder of this paper is to examine and evaluate alternative avenues for the simulationist to pursue.

4.3 Post-hoc interpretive reference-fixing

On one possible view, the immediate outputs of the ECS are in themselves referentially inert (or 'gappy') and are about events in one's past only insofar as one *interprets* them as being so.[11] Care must be taken, for there is the potential for instability here. If one generally counts as *remembering* independently of whether one has yet 'assigned' a referent, then this view collapses into the anti-referentialist view, which simulationists cannot easily accept (§3). On the other hand, if one only counts as remembering *after* one 'assigns' a referent, then we get a version of (STM-2). The view is then a claim about where or when reference-fixing occurs. Namely, not (necessarily) from the ECS, nor (necessarily) from an appropriately caused memory trace, but, rather, from a partly independent interpretative process. We must then read 'produced' in (STM-2) as 'in part produced'. And, in that case, it would make sense to talk of evaluating the *ECS-interpretation process pair* for reliability on an occasion. We may then think of this sort of view as a 'dual component' view of remembering, with the ECS responsible for simulation and some as-yet unspecified interpretative process responsible for reference-fixing.

[11] We are grateful to Denis Perrin for discussions of this view.

Unfortunately, this approach is difficult to square with the simulationist's project. If the ECS can itself be assessed for reliability, as condition (ii) in (STM) suggests, *its* outputs must (sometimes) be assessable for accuracy. And they can only be assessed for accuracy if they refer. Given the process reliabilist picture of reliability, then, *the internal goings-on of the ECS must at least sometimes determine the referent* of the episodic representation it produces. It would be difficult to maintain that the assignment of a referent to the output of the ECS is a separate task, always or generally handled by a different mechanism. If the simulationist were to explicate reliability not in terms of the proper functioning of a cognitive system but, instead, in terms of its interaction with some 'interpretative mechanism', the theory's naturalistic approach would arguably be jeopardised. First, the move feels ad hoc, against the spirit of theory's descriptive, empirically grounded underpinnings. Second, it is not easy to make sense of the notion of proper function where it is applied across systems/mechanisms or to interactions between them. Finally, if the proper functioning of a certain cognitive system does not enable us to say what remembering is or what it is for it to be reliable, the basic systems-driven methodology becomes less clear. While the post hoc approach is not incoherent, and while it would be worth asking whether alternative accounts of reliability might enable the simulationist to make the approach work (see footnote 9), that project lies beyond the scope of the present paper, and so we set the post hoc approach aside.[12]

5 Reference for simulationists: a reliabilist account

In the same way that simulationism about remembering takes its cue in part from process reliabilist views in epistemology, this section draws on theories of reference with a broadly similar heritage. Dickie's (2015) project is to answer the question: "How do the relations to ordinary things that enable us to think about them do their aboutness-fixing work?" (2017, p. 748). Dickie's proposed way of answering this question is to transcend the traditional descriptivist-causalist debate by articulating a unifying alternative that sees reference as being a matter of securing what she calls 'cognitive focus'. As Evans (1982) put it, we are among other things "gatherers, transmitters and storers of information", and these pursuits constitute "the substratum of our cognitive lives" (1982, p. 122). Dickie suggests that we have a basic need to occupy mental states that are *about* things in our environment, and that we fulfil this need by marshalling information into bodies of beliefs (or, some might say, 'mental files') that enable us to 'tune in' on objects in the world around us. This 'tuning in' or 'cognitive focus' is ultimately explicated in terms of reliability. Roughly speaking, one

[12] An anonymous reviewer suggests that this interpretive assignment of a referent could be a multi-stage, internal operation of the ECS itself. While this would not be the 'post hoc' view we discuss here, it is worth considering more generally. As the ECS is sometimes understood by simulationists (Addis 2020), the ECS lacks any proprietarily episodic source of information and may draw on schemas, etc. In that case, its determination of a referent may not be as simple as tracing the causal aetiology of a memory trace. However, we think a natural way for this approach to go is along the lines we articulate in §5. If what determines that an episodic representation refers to event *e* is the alethic reliability of these schemas, background knowledge, etc., with respect to *e*, then the view is amenable to the account we propose in §5. While there may be other reference-fixing relations these schemas and background knowledge could support, we leave it to others to formulate such an alternative.

has a body of beliefs that *refers* to an ordinary object if and only if, given the distinctive way one goes about forming the beliefs in question, there is an object the properties of which one would be unlucky to get wrong and not merely lucky to get right. More precisely, while one might make rationally blameless errors, one has a body of beliefs B about *o* if and only if *o* is the thing whose properties one would reliably get right as one went about engaging in one's B-related information-marshalling and inferential practices. Notice here that while the contents of B will typically include information acquired from causal interactions with *o*, it may also include descriptive beliefs formed otherwise.[13]

A straightforward, mechanical application of this framework to remembering would go something like this: S's current memory representation refers to event *e* if and only if *e* is the event whose properties S would be unlucky to get wrong and not merely lucky to get right, given the specific path taken by S's properly functioning ECS in producing that representation. While this is for the moment the barest outline, the picture gives us the following hierarchy of simulationist analyses, targeting remembering per se (i.e., (Q1)), followed by referential remembering (i.e., (Q2)), and finally, successful remembering. Distinguishing these different levels of analysis allows the simulationist to offer a clear and unified account of various mnemic phenomena.

S remembers if and only if

i. S now has a representation R as if of an event that was produced by S's ECS (in aiming to produce a representation of an event from S's personal past). **(Current representation condition.)**
ii. S's ECS was properly functioning and hence reliable when it produced R. **(Proper functionality condition.)**

S referentially remembers if and only if (i) and (ii) obtain and

iii. There is some event in S's personal past *e* such that *e* is the event whose features S would be unlucky to get wrong and not merely lucky to get right, given the specific path taken by S's ECS in producing R. **(Referentiality condition.)**

S successfully remembers if and only if (i), (ii), and (iii) obtain and

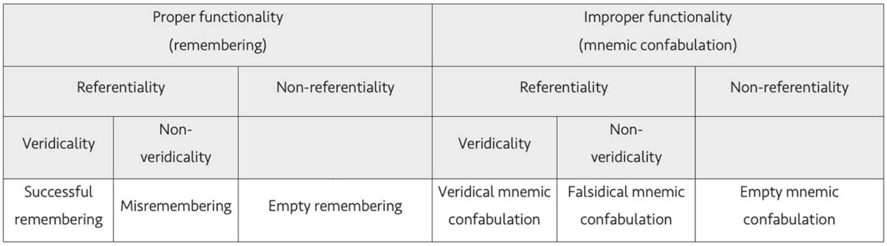
iv. R accurately represents *e*.[14] **(Veridicality condition.)**

These different levels of analysis allow us to characterise a range of mnemic phenomena, depending on which conditions are fulfilled and which are not in a given case. In Table 1 (below), the left half corresponds to cases in which both conditions (i) and (ii) are met, and the right half corresponds to cases in which condition (i) is met but

[13] As others have pointed out (Ninan 2017: 736), there is a sense in which this account of reference in terms of accuracy involves circularity, for one can only 'get an object's properties right' if one is already representing it. A subsequent debate has arisen (Dickie (2017); Openshaw (forthcoming); Pepp (2020)) as to what the explanatory ambitions of Dickie's framework may therefore be. Rather than embroil ourselves in this here, we will take up the notion of 'matching', introduced later in this sub-section.

[14] Taking fulfilment of the referentiality condition to be a pre-requisite for (in)accuracy suggests that we re-classify some phenomena. For example, falsidical 'lost in the mall' cases have sometimes been thought of as misrememberings, where here they would likely fall under 'empty remembering'. So-called 'lucky veridical lost in the mall' cases, due to the significant presence of luck, would also likely fail to fulfil the referentiality condition. (See Michaelian (2023: 137–139) for discussion of the cases in question.).

Table 1 An illustration of the various mnemic phenomena (bottom row) characterised by permuting which of conditions (ii), (iii), and (iv) are satisfied



Proper functionality (remembering)			Improper functionality (mnemic confabulation)		
Referentiality		Non-referentiality	Referentiality		Non-referentiality
Veridicality	Non-veridicality		Veridicality	Non-veridicality	
Successful remembering	Misremembering	Empty remembering	Veridical mnemic confabulation	Falsidical mnemic confabulation	Empty mnemic confabulation

condition (ii) is not. The second row then divides up these cases into those in which (iii) is met and those in which (iii) is not met. Finally, the third row divides up *those* cases into those in which (iv) is met and those in which (iv) is not met.[15]

Cases of *successful remembering* are good cases: the subject's ECS is properly functioning and hence reliable when it produces representation R, R successfully refers to some past event *e*, and R accurately represents *e*. In cases of *misremembering*, although the subject's ECS is properly functioning and hence reliable when it produces R, and although R successfully refers to some past event *e*, R inaccurately represents *e*. In the bottom right corner, corresponding to cases of *empty mnemic confabulation*, everything goes wrong.

The remaining possibilities of *empty remembering*, *veridical mnemic confabulation*, and *falsidical mnemic confabulation* are particularly interesting. If they are to be genuine rather than merely conceptual possibilities, the two notions of reliability—that appearing in (ii) and that appearing in (iii)—should in principle come apart in both directions. In other words, it must be that the ECS can properly function on an occasion despite producing a representation that fails on that occasion to refer, and that the ECS can improperly function on an occasion despite producing a representation that succeeds on that occasion to refer. So, in empty remembering, the subject's ECS is properly functioning and hence reliable when it produces representation R, but R fails to in fact pick out any past event and hence cannot be evaluated for accuracy. Equally, in veridical and falsidical mnemic confabulation, the subject's ECS is improperly functioning, or in any case is unreliable when it produces R, but nevertheless R manages to lock on to some past event and either (in)accurately characterize it. We will say more to explain how the two notions of reliability can come apart to produce these results below, once we have articulated condition (iii) with more precision. Cases of empty remembering are discussed in §5.2. Cases of referential confabulation are discussed in §6.[16]

[15] It is interesting to consider how this system of classification interacts with others that have been proposed. One may think of it as a refinement of Michaelian's (2016a) table, where the phenomena of successful remembering, misremembering, and veridical/falsidical confabulation are now restricted to cases where the referentiality condition is fulfilled. Illustrating how this system interacts with others, including those that feature a meta-level component (Michaelian 2023), is left for another occasion.

[16] To keep discussion manageable in this paper, we mostly set aside exactly how the simulationist precisification of (ii) and (iv) ought to go. These are broader questions than we are able to address here.

To postpone circularity worries concerning the account of reference in terms of accuracy, which in turn presupposes reference, we will understand the notion of 'getting an event's properties right' in the Referentiality Condition above in terms of *matching* rather than accuracy per se. To take a basic example, matching an event e's properties is a matter of possessing an event-representation produced by the ECS that attributes F-ness (e.g., redness at a certain location) and there being some event e which instantiates F. There will, no doubt, be many events that 'match' any given event-representation. This is where the role of the fine-grained paths to producing an event-representation enter in, along with the safety-like modal tracking requirement: actual matching is not enough. But such details are, for now, left for presentation in §5.2.[17]

We will be assuming that accuracy (and therefore matching) in remembering is generally a matter of one's episodic representation approximating the features of mind-independent events in one's personal past rather than of one's particular experiences of those events. This *alethist* (Michaelian & Sant'Anna, 2022) as opposed to *authenticist* (Bernecker, 2015; McCarroll, 2018) assumption helps to simplify our presentation. But it is also a natural partner for postcausal—and, in particular, simulationist—theories of remembering. According to authenticism, one accurately remembers an event as having been F only if, at the time of one's experience of the event, one experienced it as being F. Given the vast psychological literature on memory distortions, the apparent regularity with which they occur, and the adaptively beneficial content-modulating mechanisms they seem to indicate, to embrace authenticism is, at least superficially, at odds with the simulationist's project.[18]

5.1 Reliability without appropriate causation

Before we get on to discussing ways to precisify the framework above, and in particular condition (iii), we first want to address a knee-jerk complaint: *What could underpin such reliability, if not causal chains?* Michaelian's (2016b) claim is not that remembering never involves a causal connection to the past event (perhaps even often an 'appropriate' one). It just isn't necessarily so. The referentiality condition sketched here in §5 can be fulfilled by appropriate causation in some cases even if it isn't secured in that way in all cases. Before sharpening the referentiality condition, we illustrate one way for an output of the ECS (when it is properly functioning) that is a candidate for referring to a particular event might succeed in doing so by virtue of fulfilling something like the referentiality condition above despite failing to involve an appropriate causal connection to that past event.

Suppose there is an experienced event in your personal past, e. And suppose that, since its occurrence, you have lost any privileged memory trace that might have once afforded an appropriate causal connection to e. In trying to piece together what must have happened that day, in part on the basis of related things you can remember in the causalist's way, items of semantic memory, and generally reliable patterns of inference,

[17] Thanks to an anonymous reviewer for prompting us to clarify this.

[18] Whether authenticism really is as difficult to reconcile with the constructive picture of memory presented by psychology as has been supposed is debatable (Openshaw, 2023).

you start to reconstruct what must have occurred. Think of yourself as piecing together a number of building blocks corresponding to the place and time of event, the people involved in the event, etc. These building blocks are not traces but more like object concepts, schemas, or general models (Ghosh & Gilboa, 2014). They may involve a kind of reference to particular places and individuals, but their referentiality is not to be explained in terms of appropriately discriminating causal chains to one-off encounters: there is a sense in which these models get richer through repeated exposures. It is not implausible to think that the referential characteristics of these building blocks could serve to considerably narrow down the candidate referents of episodic simulations in which they participate. Though impressionistic for the moment, one can begin to get a sense for how the referentiality condition sketched above could be fulfilled with respect to some event *e* in a case of remembering lacking in any appropriate causal connection to *e:* one could certainly in principle have a reasonable degree of reliability (in a sense to be refined) even if there is nothing uniquely discriminating about the causal aetiology of the information that is encoded in the various elements or building blocks of the construction process. In principle, many of the building blocks may have been acquired by means other than first-hand experience. Nevertheless, there could still be a particular past event with respect to which one fulfils the referentiality condition.[19]

To help make this suggestion more concrete, consider an example.[20] Suppose you're trying to remember a conversation your notes indicate that you had last year with a colleague at the office. Immersed in philosophical dialogue, you paid relatively little attention to your usual surroundings. Were you to now try to recall details unique to that occasion—whether so-and-so's office door was open or closed—you would likely fail, and any success would be mere luck. But were you to remember the many details constant in that environment, you would do so with a high degree of reliable accuracy, thanks to your trusty cognitive map and items of semantic memory such as schemas for the office and how things there are laid out in space or generally look, your background knowledge about how the colleague in question tends to act or think, scripts for how conversations tend to go, etc. (Binder & Desai, 2011). This is not just an armchair intuition pump. As Addis (2020) puts it, "[a]lthough much remains to be determined with respect to the role of schema in event simulations, what is clear is that schemas are essential to their construction" (p. 246). Many contributions to the psychology and neuroscience of memory turn on the assumption that when semantic

[19] 'Problem of the many'-style worries may seem to raise a special threat to this sort of account. Given the abundance of metaphysically precise time slices, for any particular event there will be many almost exactly overlapping distinct events. A putative rememberer who reliably gets event *e*'s properties right will also, then, reliably get the properties of many other events properties right, differing only by nanoseconds. Of course, this kind of underdetermination worry also poses a threat to (CTM), since there will also be no metaphysically precise time corresponding to the terminus of an appropriate causal link. While we agree this is a delicate issue, it is perhaps inevitable that, strictly speaking, we never remember just one metaphysically precise event. In the same way, we never refer to a metaphysically precise location when we demonstratively refer to a location as 'there'. A useful resource for anyone here is the notion of 'multiple reference': in many cases where there is a tie for the unique most eligible referent, we can conclude that the relevant representation bears multiple reference relations to many distinct things (Openshaw 2021). Thanks to Ali Boyle for emphasising this point to us.

[20] A similar case is described by Andonovski (2022, pp. 12–15).

memory is used to supplement remembering in this sort of way, this does not mean that the elements contributed by those means are not ‘really remembered’ but only ‘inferred’, ‘known’, or ‘imagined’. Such distinctions are simply absent from these studies. Whether accurately remembered information is derived from the retrieval of a memory trace originating in that event or from semantic memory built up from experiences of congruent events is irrelevant to their research (see, e.g., Fayyaz et al. (2022) and Zöllner et al. (2023)). And for good reason: normal, everyday remembering is not the activity of an episodic memory system in isolation. Episodic and semantic memory “are inextricably intertwined” (Renoult et al., 2019) and their neural correlates largely overlap (Binder et al., 2009). Psychologists increasingly warn against obscuring “fundamental interdependencies and indeed, gradients” between episodic and semantic memory (Strikwerda-Brown et al., 2022, p. 618). Given all of this, we think it is very plausible that causalists overestimate how much of the burden of reference-determination is carried by memory traces. As a result, they underestimate the likelihood that cases like that just considered are more typical than we would give them credit for if we denied that such cases involved genuine mnemic reference. So, when you remember the conversation with your colleague last year, it probably matters much less than philosophers have sometimes thought whether your event representation accurately matches the event by virtue of the retrieval of a memory trace originating in your experience of the event or, instead, by virtue of input from items of semantic memory that have a less discriminating causal aetiology, being built up and abstracted from multiple experiences of relevantly similar events.

Before precisifying the referentiality condition, we pause to note an additional virtue of the reliabilist approach to reference-fixing: this sort of account promises to also extend to reference-fixing in cases of episodic future thought.[21]

In 1512, Henry VIII ordered the construction of a warship: the *Henry Grace à Dieu*.[22] It was to be 50 m long and have a forecastle 4 decks high. Imagine Henry knows his order will be carried out on time but is given no reports of progress during its construction (on the principle that no news is good news). A short time after its expected completion, he sets off towards the Thames where the ship awaits. As he does so, he has various thoughts about *Henry Grace à Dieu*, such as he expresses when self-satisfyingly proclaiming ‘*Henry Grace à Dieu* is the largest warship in Europe’. Some of these thoughts may be episodic future thoughts, such as when he prospectively imagines *seeing that ship* for the first time.

In this case, the causal relation runs in the wrong direction to conform with the usual causalist account. Insofar as Henry has beliefs about the ship, they are not causally derived from it. Before indicating how the referentiality condition can explain this sort of case, we can motivate it further by also considering a modified version. Suppose that things are just as in the vignette above, only the builders of the sea vessel, in defiance, have constructed a small, unarmed yacht meeting none of the king’s needs. In this case, Henry VIII’s thoughts as he travels towards the river are intuitively *not* about the yacht. A second insight, then, is that whatever relation there might indeed

[21] This may be an appealing virtue for continuists about the relation between remembering and constructive imagination (see footnote 8).

[22] This case is adapted from Hawthorne and Manley (2012: 28). For similar cases, see Jeshion (2010).

be between King Henry and the vessel, via his men, it is *insufficient* to determine the object of Henry's thoughts if it does not enable him to reliably identify its properties.

We can explain each of these cases and their respective asymmetry in securing reference as follows. In the first case, although Henry's beliefs do not causally derive from encounters with the ship, they were formed by a means which reliably gets the ship's properties right, at least for those properties with respect to which Henry has, or is disposed to form, beliefs. And so the reliabilist view, albeit inchoate for the moment, accommodates the first case in an appealing way: Henry's means of constructing an episodic representation in which he first sets eyes on the ship *suffices* to put him in a position to refer to the *Henry Grace à Dieu*. What is more, in the modified case, the reliabilist view suggests that part of what is *necessary* for successful reference is the presence of a means by which the subject can reliably identify the event's salient properties. It is because this is absent in the modified case that Henry VIII is incapable of constructing an episodic representation which successfully refers to the *Henry Grace à Dieu*.

The reliabilist view affords the following general insight: the presence of a causal link running from object to thinker is typically involved in securing reference *because it is typically involved in their having a means of forming representations that reliably get its properties right*.[23] But this fact about how we typically put ourselves in a position to form accurate representations should not be mistaken for a necessary condition.

It suffices for us to provide motivation for the reliabilist view that it can accommodate Henry's successful episodic future thought in the original *Henry Grace à Dieu* case, for it is not at all clear what the causalist can say here. Yet, in presenting this material, we have encountered some resistance to the account's verdict in the modified *Henry Grace à Dieu* case, in which Henry's episodic representation apparently fails to refer to the small, unarmed yacht.[24] We do not deny there is some temptation to think Henry's episodic future thoughts refer to the yacht and that he simply misrepresents it. However, if singular thought constitutes a genuine cognitive success, it cannot be had too easily. Consider, by contrast, that according to some it is enough to have a singular thought about a particular thing if one introduces a name with the stipulation that it refer to some unique F. For example, one introduces the descriptive name 'Newman1' by stipulating that it refers, if it refers at all, to the first child to be born in the twenty-second century (Kaplan, 1969). According to some 'liberal' views, one

[23] Close followers of the Kripkean tradition would reject this suggested link between successful reference and any kind of reliability concerning the purported referent's properties. For them, what would matter is roughly just that the right kind of causal relation obtains, not whether the subject is in a position to make effective use of it. Dickie (2015, 157ff) makes an interesting case, however, that causalists ought to not only point to a type of relation which in fact obtains in all cases of successful reference, but to explain *why that* relation plays an aboutness-fixing role. In doing so, we think that something like the framework Dickie proposes—and on which we draw here—becomes attractive. Thanks to an anonymous reviewer for asking that we clarify this point.

[24] It could be argued that even in this modified case there is a good enough link of causal or counterfactual dependence between Henry's intentions and the yacht, for without those intentions nothing would have been constructed. Our concern is that it matters to some extent what the men build for the link to be good enough. Had they assembled all of the ship parts to construct a bizarre tower, it would still be true that without Henry's command nothing would have been built. But that would not be enough to show that Henry's thoughts, apparently about a warship, were really about a tower. We would like to thank an anonymous reviewer for pressing us to say more about this case.

is thereby afforded the capacity to have singular thoughts about Newman1 of the sort which would be expressed by the use of that name (Hawthorne & Manley, 2012). In contrast, we think of singular thought as a genuine cognitive achievement; one cannot "produce new thoughts […] by a 'stroke of the pen'" (Evans, 1982, p. 50). As Kaplan (1989) suggested, although a language with the name 'Newman1' "enables us to *express* contents that would otherwise be inaccessible […], something more, something like being *en rapport* with the [individual], is required to *apprehend* the content (and thus to hold attitudes toward it)" (pp. 606–607). Similarly, we think that it is not enough to have genuinely referential imaginings about a particular if one simply gives it a name, or opens a 'mental file', if there is not also some form of reliable epistemic dependence of the subject's mental states upon it. If Henry lacks such a relationship with the yacht (as we are supposing in the modified scenario), then although one might *interpret* or *describe* him as intending to think about it, and although his utterances of sentences containing the name '*Henry Grace à Dieu*' arguably refer to the yacht, this shouldn't be taken to show that he is capable of genuinely referential thinking about it if he has no reliable means of representing it in the relevant sense. If reference at the cognitive level were so easy, the very significance of the distinction between singular and descriptive intentionality would be jeopardised. In sum, we think that the reliabilist account has the virtue of being able to account for reference-fixing in cases of episodic future thought without resulting in a kind of eliminativism on which the distinction between singular and descriptive intentionality fails to track any real cognitive difference.

5.2 Precisifying the referentiality condition

We will now propose a natural way of precisifying condition (iii), and in particular the role played by the informal notion of 'luck'.

The 'not merely lucky to get right' component captures the sense in which accurately *remembering* an event has, in the normal course of things, a reasonably robust explanation. If one is remembering an event and one does succeed in getting its features correct, this is not, in the normal course of things, due to blind chance. While this need not be due to appropriate causal connection (§5.1), reference goes hand-in-hand with a general tendency to 'match' a thing's properties. The 'not merely lucky to get right' clause can be sharpened as follows:

there is some event e for which it is true that, given the specific path p taken by S's ECS in constructing representation R on this occasion, and given the range of event-features R represents, $f_1 \ldots f_n$, R would not easily attribute feature f_i if e did not possess feature f_i.

The idea is that which event one is remembering is a matter of whether there is some event (or are some events) whose salient properties one's event-representation matches not only at the actual world but also at worlds that are suitably close to the actual world.[25]

[25] There is a structural similarity between the question of how to individuate what we are calling 'paths' and the generality problem for reliabilism. See Grundmann (2018) and Tolly (2021) for recent discussions.

If the sharpening in the previous paragraph were left unsupplemented, it would follow that for every event-feature represented by R, the referent of R has that feature. Given that R attributes feature f_i and that 'R would not easily attribute feature f_i if *e* did not possess feature f_i', then the referent of R must have instantiated feature f_i. This corollary would make referential remembering extremely demanding. Misremembering the colour of the jumper one's friend was wearing at the restaurant last week would be incompatible with remembering that evening's dinner. So there must be some restriction on *which* represented event-features are relevant to reference determination.

We think that it is extremely unlikely that there is any *particular* feature the incorrect attribution of which, in every case, leads to reference-failure. In part this is because of the great variety in the possible objects of remembering, but also due to the plausible idea that incorrectness in one feature (e.g., the colour of a friend's jumper) can almost always be outweighed by correctness in others. Moreover, which features are salient intuitively varies across situations, and this is a feature that mnemic reference arguably has in common with reference more generally (Dickie, 2015, 175ff). The idea is that the referentiality condition need not be satisfied *for every* feature attributed by R. Referential remembering requires only that one *generally* get the event's properties right.

Precisified referentiality condition there is some event *e* for which it is true that, given the specific path *p* taken by S's ECS in constructing representation R on this occasion, *for a general range* of the event-features R represents, $f_1 \ldots f_n$, R would not easily attribute feature f_i if *e* did not possess feature f_i.

To begin with a very basic example, in entirely ordinary circumstances Sally (a neurotypical subject with well-functioning memory capacities) remembers having dinner with her friend Alice last week but incorrectly represents her as having worn the blue jumper she very often wears in winter when in fact she was wearing a green jumper. We know that one of the functions underpinning the constructive character of her ECS on this occasion—e.g., its reliance on semantic memory, schemas, and general knowledge—is to safeguard accuracy at an efficient cost (e.g., Aronowitz (2019)). Moreover, the colour of her friend's jumper is unlikely to be part of the autobiographical significance of the event for Sally, nor is it likely to be practically significant to her present interests. In this case, the proper functionality and (precisified) referentiality conditions are met, but the veridicality condition is not met, and so it is a simple case of misremembering.[26]

Next consider an instance of *empty remembering*, where the proper functionality condition is met but the (precisified) referentiality condition is not met. Suppose that Alice is prompted to recall having dinner with her friend Sally last week but situates her representation as of the event in the wrong restaurant, or even in the wrong city. Where they dined *is* likely to be part of the autobiographical significance of the event for Alice,

[26] Similar referentially innocuous memory errors might include, for example, the so-called visual Mandela effect (Prasad & Bainbridge, forthcoming) and event completion (Strickland & Keil 2011).

and it is also likely to be practically significant to her present interests. In this case, the proper functionality condition is met, but the (precisified) referentiality condition is not (nor, *a fortiori*, is the veridicality condition). While it might be tempting, as an outside observer, to *interpret* Alice as remembering the particular event, albeit unsuccessfully, we should not conclude from this that Alice *is* in fact managing to refer to the event while characterising it inaccurately. As Evans (1982, 130ff) notes, correctly interpreting someone as, in some sense, having intended to think about x does not entail that they were in a position to think about x.

Talk of 'generally' getting an event's properties right is not particularly precise, and this is likely to strike some as a serious weakness. We will say a little more in the remainder of this sub-section to articulate the kinds of (in)accuracy that tend to weigh strongly as well as the kinds of situation-dependent parameters that can have an effect on what dimensions of (in)accuracy are referentially significant on a given occasion. First, though, we should clarify our methodology and the scope of our ambitions.

Our goal here is articulate a framework—and to indicate its benefits—for post-causalists (and simulationists in particular) to begin thinking more seriously about reference in remembering. It should not be a constraint on the value of this account that it makes predictions about whether reference in achieved in difficult cases. Compare: we do not ask that reliabilist conditions on knowledge determine how many miles away Henry must be from a fake barn to count as knowing that what he sees is a barn.[27] The explanatory utility of safety principles in epistemology do not carry a commitment to specifying, in advance, exactly what similarities are tracked by the notion of 'closeness'.[28] Nor do we ask that possible worlds semantics for counterfactuals say much to determine how 'distant' a not-q world must be if a given counterfactual is to be true. As Lewis (1986, pp. 91–5) put it, the vagueness of the account matches the vagueness of our judgments (see also Williamson (2009, pp. 9–10)). "While not devoid of testable content [...] it does little to predict the truth values of particular counterfactuals in particular contexts" (Lewis, 1979, p. 465). The important point is that this is clearly no defect of the account's utility. Lewis suggests that in supplementing the account with a specific characterisation of the similarity relation, there is ultimately nowhere else to fall other than on our intuitive judgments about particular cases. The account itself is not designed to make testable predictions which we may then verify. Analogously, our account of the conditions under which there is sufficient reliability to secure reference in remembering ought to leave available a variety of precisifications for settling more recherché cases—precisifications to be whittled down as our overarching theory of memory progresses. Much like the theoretical utility of modal epistemology and of possible worlds semantics, the utility of a reliabilist picture of mnemic reference-fixing does not critically hang on its ability to specify clear and tidy predictions about difficult cases in advance of its application to theorising.'[29,30]

[27] We are alluding to the case described by Goldman (1976).

[28] As Hawthorne (2003: 56, n. 17) emphasises, we also cannot rely on any general-purpose notion of closeness at play in discussions of counterfactuals.

[29] It is also worth noting that our competitor's talk of 'appropriate causal relations' is little more precise.

[30] In some cases, it is also not clear what the right verdict is. Suppose Aya recently taught a 3-hour class in room R. Call this extended event *e1*. Since she accidentally left behind a USB drive, she briefly stopped

At any rate, we can say some more systematic things about our situation-sensitive notion of general accuracy in the precisified referentiality condition.

It is widely agreed that the constructive character of episodic memory is not so much a defect as, in part, a reflection of how remembering serves the specific goals of the rememberer in their present social and practical contexts (Neisser, 1997). What counts as 'good' or 'successful' remembering "often involves getting something right about the significance of the past as judged from the standpoint of the present" (Campbell, 2006, p. 362). For Campbell (2006), part of what this means is that remembering has multiple dimensions of 'goodness' (i.e., what she calls integrity), extending beyond mere accuracy. Another, however, and more pertinent to our discussion, is that *which* features of an event are relevant for evaluating accuracy varies with the rememberer's present social and practical concerns. Given the connection between accuracy and reference in the precisified referentiality condition, it follows that the conditions for successful reference might also vary with changes in the rememberer's concerns. But although we should be alert to the situation-sensitivity of the precisified referentiality condition, we can nevertheless make some general observations about the dimensions of matching that typically matter and how these can affect the successfulness of mnemic reference.

First, *spatiotemporal* features typically carry a good deal of weight. Misremembering the location or the general time period in which the event took place can be compensated for if one accurately remembers sufficiently many other salient features. But, in general, and as illustrated by the case above in which Alice is prompted to remember having dinner with her friend Sally last week, a failure to match in these respects can be enough to produce a failure of reference.

Second, autobiographically or practically significant features typically carry more weight than other kinds of features. If one systematically fails to achieve matching with respect to *who* was present at the time of some social event (e.g., that one met Sarah for lunch rather than Sally), or one's actions during the course of the event, this might also contribute to generating reference-failure.

Third, metarepresentational features concerning the *source* of the first-order event-feature representations count strongly in the balance. If one remembers a dreamt, imagined, or vicariously experienced event *as* an event veridically perceived first-hand,

Footnote 30 continued
by to pick it up the following day. Call this brief episode *e2*. For whatever reason, Aya has since lost any specific trace with its origins in *e2*. Yet suppose she is remembering *e2* via the kind of 'building block model', schema-based means we described in §5.1. Given how this model was formed, it will more reliably match *e1* than *e2*. Can the reliabilist account we propose here accommodate this case? Or will it turn out that Aya is actually misremembering *e1* rather than accurately remembering *e2?*We note first that such cases are as much a problem for causalists. Second, given the different dimensions of matching that affect reference-determination, we are inclined to reply that the autobiographically or practically *salient* properties in this case include Aya's actions; in particular, whether the event features actions such as retrieving a USB drive and dashing off, or lecturing and interacting with students. If such properties are more highly weighted when it comes to reference-determination, we can expect the right result. This variable weighting of event features for the purposes of reference-determination enables our account to avoid much of the underdetermination that might otherwise often result from the less discriminating causal aetiology of items of semantic memory in contrast to the causalist's monogamous traces. Finally, however, it is also worth noting that not all intuitions are likely to be respected by the overall best theory of reference in remembering, and given the theory-laden set-up involving memory schemas, it is not obvious we have clear cut, 'untarnished' intuitions here. We thank an anonymous reviewer for bringing this sort of case to our attention.

it is natural to think this kind of matching failure often results in more than simply misremembering the event but in failing to remember *it* at all. Source and reality monitoring processes at the time of remembering capitalize on average differences in the quantity and quality of perceptual, affective, spatial, and semantic information across memories from different sources (Johnson et al., 1993). While memories for imagined versus perceived events tend to include fewer sensory or spatial details, and to include more information about cognitive operations (Ibid.), subjects nevertheless sometimes mistake previously imagined for previously perceived events (Gonsalves et al., 2004; Henkel et al., 2000). When they do, it is plausible that this can lead to reference-failure.

In sum, and substituting in the precisified referentiality condition, we end up with the following unified, multi-level account:

S <u>remembers</u> if and only if

i. S now has a representation R as if of an event that was produced by S's ECS (in aiming to produce a representation of an event from S's personal past). **(Current representation condition.)**
ii. S's ECS was properly functioning and hence reliable when it produced R. **(Proper functionality condition.)**

S <u>referentially remembers</u> if and only if (i) and (ii) obtain and

iii. There is some event *e* for which it is true that, given the specific path *p* taken by S's ECS in constructing representation R on this occasion, *for a general range* of the event-features R represents, $f_1 \ldots f_n$, R would not easily attribute feature f_i if *e* did not possess feature f_i.**(Precisified referentiality condition.)**

S <u>successfully remembers</u> if and only if (i), (ii), and (iii) obtain and

iv. R accurately represents *e*. **(Veridicality condition.)**

One might query whether this uniform, reliabilist account of reference-fixing is, in the end, meaningfully different from the pluralist view discussed in §4.1.[31] On that view, the reference-fixing conditions for remembering are situation-dependent. Appropriate causation via a memory trace may be what determines the referent of an occurrence of remembering in many situations, but in others the referent of an occurrence of remembering will be determined by a different means. The most that can be said about reference-fixing in general, on this view, is: it depends. The main difference between pluralism and the reliabilist account endorsed here is that reliabilists can explain reference uniformly. There is just one condition to be met for successful reference (i.e., the precisified reliability condition). No doubt this condition is itself situation-dependent. But the reason this explanatory uniformity (of reference in terms of reliability, rather than a plurality of conditions) is important is not just for book-keeping purposes. The reason is that this reflects a more general picture of what the explanatory basis of cognitive reference is. The pluralist gives us no indication as to *why* their diverse set of conditions are reference-fixing. If one were to ask the pluralist why condition *c* is enough to achieve reference in scenario *s* but condition *c'* is not

[31] Thanks to an anonymous reviewer for inviting us to consider this issue.

enough to achieve reference in scenario *s'*, it is not clear they could say anything at all. In contrast, the reliabilist proposal furnishes us with a unified explanation of *why* this or that feature of scenario *s* secures reference. It can also tell the *same* story for future-directed imagination and other cognitive capacities. This is because cognitive reference, or singular thought, is in general a relation one has to a thing (roughly) in virtue of it being that which one is disposed to accurately represent at relevant nearby worlds. Whether that broad conception of the explanatory basis of reference is true or not is, of course, a big, open question. But we think that it is an explanatory virtue of the reliabilist account endorsed here that it has something to say in the face of such questions, and on these pluralism is regrettably silent.

6 Referential mnemic confabulation

In this section, we move from motivating and outlining our account (§5) to exploring some of its implications. Perhaps the most interesting conceptual possibilities illustrated by Table 1 are cases of veridical or falsidical confabulation. These will be cases in which a subject has a representation R as if of an event but which was produced by an *improperly functioning and hence unreliable* ECS. Nevertheless, they are also cases in which R succeeds in referring to some event *e* and is either accurate with respect to *e* or inaccurate with respect to *e*. In other words, they are cases in which conditions (i) and (iii) are met but (*unlike in referential remembering)* condition (ii) is not. These cases will be possible only insofar as it is possible for the ECS to be functioning unreliably on an occasion whilst *reliably* getting some event's properties right on that same occasion. Are such cases possible?[32]

As a preliminary, it must be clarified that the term 'confabulation' has been used in many different ways by theorists with different interests at different times. Over fifty years ago, Berlyne (1972) lamented that the term was "widely employed, poorly defined and variously interpreted" (31), and its scope has arguably continued to expand. Whereas some use the term to refer only to pathological errors observed in clinical contexts, others extend the term beyond such contexts to relatively mundane errors. Yet others apply the term to phenomena not all of which are best described as *memory* errors at all, for example anosognosia (as Robins (2019, p. 2140) observes). Even in cases which do crucially involve defects of memory—we follow Bernecker (2017) in calling these *mnemic confabulations*—it is useful to draw a distinction between what Bortolotti and Cox (2009, p. 954) call *primary* versus *secondary* confabulations. Whereas primary confabulations will then be a certain sort of defective product of the memory system, secondary confabulations are a derivative phenomenon, wherein the sufferer of a primary confabulation produces claims in an effort to justify their initial confabulation in response to challenge.[33] Finally, confabulations differ in their

[32] This question may also depend on whether there are *other* functions the ECS might fulfil reliably on an occasion whilst failing to reliably fulfil the function of representing some particular event(s).

[33] For example, a patient who reported having been married for only 4 months was pressed to explain how they had come to have four adult children. In response the patient claimed (falsely) that the children had been adopted (Moscovitch 1989: 135–6).

formal features, and it is worth distinguishing the infamous cases of *fantastic* confabulation, which “have no basis in reality and are intrinsically nonsensical and illogical” (Schnider, 2018, p. 68), and those which are more mundane and superficially sensible. The utility of this distinction is compatible with the claim that they result from the failure of control processes of the same kind (Burgess & Shallice, 1996). Despite the diversity of phenomena to which the term has been applied, some have proposed unifying theories of confabulation understood in this broad way (e.g., Hirstein (2005)).

In what follows, we exclusively have in mind (primary) mnemic confabulations, in particular those which are episodic, past-oriented, and which need not be fantastic or occur in clinical settings.[34]

While there is no exhaustive and universally accepted definition of mnemic confabulation, a useful approximation describes it as a symptom that

> consists in involuntary and unconscious production of ‘false memories’, that is the recollection of episodes, which never actually happened, or which occurred in a different temporal-spatial context to that being referred to by the patient (Dalla Barba, 2002, p. 28).

Crucially, it is widely assumed that such confabulations are produced by impaired memory processes and *not* by distinct, compensatory procedures (Burgess & Shallice, 1996, p. 361).

Schnider (2018) describes a patient known as Mrs. B, who “confabulated events that had not taken place, falsely recognized people, confused the day and the place, and confabulated obligations that she did not have at the present time”, but most of whose confabulatory representations “referred to real events and experiences in her past” (2018, p. 7). Often, when mnemic confabulations do purport to concern the subject’s personal past, distortions are a matter of ‘erroneous temporal reference’ (Schnider, 2018, p. 205). Berlyne’s (1972) suggested distinction between momentary and fantastic confabulations is that “[t]he former are temporally displaced true memories, the latter wish-fulfilling fantasies” (p. 38). Indeed, this has sometimes been thought a key feature of confabulatory phenomena:

> The source [of confabulations is] predominantly […] the patient’s actual experiences in an earlier phase of his life. [Confabulation] seems to arise from the disruption of his temporal frame of reference, so that true statements become displaced in their chronological setting, those drawn from different periods become confused. Typically, a memory of their more remote past re-emerges as an event in the present or immediate past (Talland, 1965, p. 56).

In the remainder of this section, we will tease apart the sense in which mnemic confabulations are products of an improperly functioning and hence unreliable episodic memory system from the claim that they fail to refer to events in the confabulator’s personal past. We will argue that this divorce between mnemic reference and genuine remembering is best accommodated by our account (§5).

[34] See Robins (2020) for a discussion of the relation between broad and mnemic confabulation. Not all confabulations are episodic in character (Baddeley and Wilson 1986; Moscovitch 1995), nor are all about events in the subject’s personal past as opposed to their future (Dalla Barba 2009).

We assume that mnemic confabulations are the result of an improperly functioning memory system (Michaelian (2016b, p. 109; 2021, p. 7491)). In many cases of confabulation, a subject correctly remembers enough about some particular event(s) in their personal past but, because they dramatically mislocate the event within the chronology of their life, they do not count as genuinely remembering. Nevertheless, on that occasion, their ECS produces an episodic representation which succeeds in referring to the event(s) in question.

In Dalla Barba et al.'s (1990) description of CA, a Wernicke-Korsakoff Syndrome patient, identified one set of responses which, unlike the more bizarre responses CA would occasionally give to other prompts, involve the "recall of a real past event in the wrong temporal context", such as (when prompted to recall what she did last Christmas): "I went to church and came back home to help my mother cooking Christmas lunch for me and my brothers", although "[h]er mother actually died 15 years previously and one of her two brothers 30 years previously (p. 530).

It seems perfectly coherent for there to be instances of confabulation which are nonetheless *about particular events in the subject's past*. The commentary in Dalla Barba et al.'s (1990) study suggests that CA was referring to general or repeated events from her past when she claimed to have celebrated last Christmas with her mother and two brothers, or to have met with salesmen and clients earlier in the day. A natural thing to say here is that CA is confabulating but is nevertheless successfully picking out events from her personal past. Depending on just how accurate CA's episodic representation of the past event(s) in question really was, it might be that there actually are genuine instances of veridical or falsidical confabulation. That is, it may be that CA's ECS was improperly functioning—or was in any case unreliable at a system level—even though the specific process carried out on that occasion achieves a kind of focus on some event(s) in her past. In sum, there is conceptual room for this possibility, and some forms of confabulatory phenomena from the empirical literature may provide us with actual instances.

The phenomenon of referential confabulation offers a point in favour of the kind of theory we presented in §5 and against many causal theories of remembering. To illustrate, consider Robins' (2020) account of mnemic confabulation, according to which it occurs "when there is no relation between a person's seeming to remember a particular event or experience and any event or experience from their past—either because there is no such event in their past or because any similarity to such an event is entirely coincidental" (125–6). Similarly, on Bernecker's account, mnemic confabulations differ from instances of genuine remembering in that "they fail to be suitably causally connected to the corresponding past representations, either because there are no corresponding past representations or because the causal connection has been severed" (p. 12). In other words, "[i]t is the hallmark of confabulation that any match between the contents of the past and present representations is nothing but a lucky accident" (2017, p. 9).[35]

The reason that these theorists characterise mnemic confabulation as constitutively failing to be about particular events in the confabulator's personal past—and therefore

[35] Werning & Liefke (forthcoming) also claim that mnemic confabulations constitutively lack referents.

as necessarily ruling out the possibility of *referential* confabulation—is their background endorsement of a causal theory of remembering. For, from this starting point, if it were granted that the confabulator has an apparent memory representation that *refers* to a particular past event, it would follow that their representation is appropriately caused (by means of an memory trace), and it would *therefore* follow that they are simply *remembering*—not confabulating! Since, for causalists, it is sufficient for one to be remembering an event that one have an apparent memory representation that is appropriately caused, via a memory trace, by one's past experience of the event, no *confabulation* can have this same feature. Confabulations must "lack [such] a connection to any event in the confabulator's past" (Robins, 2020, p. 130); they are "errors because they lack a causal connection between the event and its representation" (Robins, 2020, p. 126). In order to predict and explain cases of referential confabulation, we need to prise apart our account of mnemic reference and our account of remembering. And this is precisely what the theory proposed in §5 enables us to do.

Of course, causalists could give a separate account of referential confabulation. But we think it is a clear advantage of the framework proposed here that it predicts the existence of these hitherto neglected cases and provides a uniform, non-ad hoc account: referential confabulations are products of improperly functioning episodic memory systems that nonetheless, on that occasion, produce a representation that fulfils our (precisified) referentiality condition (§5.2).[36]

We do not envision that this is anything like a knock-down objection, and indeed we acknowledge that the causalist has at least two lines of reply available to them. First, they might insist that what we are calling cases of referential confabulation are merely cases of misremembering. For example, C.A. is remembering some past occasion(s) of cooking Christmas lunch with her mother but misremembering when it happened. In that case, the causalist can deploy their usual account of mnemic reference. Second, the causalist could instead suggest that referential confabulation is *compatible* with misremembering. In particular, C.A. is indeed confabulating, but she is also misremembering, and it is in accounting for the latter fact that the causalist can deploy their usual account of mnemic reference. Without anticipating at length the delicate issues that this dialectic raises, we would like to indicate that neither reply is satisfying.[37] In response to the first reply, the psychological literature does not characterise these cases as mere misrememberings, and (intuitively) being 'lost in time' is a graver error than this categorisation suggests. Genuine (mis)remembering involves proper retrieval, and retrieval requires more than that some trace or other drive an instance of episodic simulation. The right trace must be identified by proper search mechanisms and its output properly monitored and evaluated against the cue demands (Burgess & Shallice, 1996). Moreover, given just how frequently confabulation is tied to temporal displacement in the psychological literature, this reply threatens to dictate,

[36] Alternatively, the causalist could offer separate accounts of the kind of appropriate causation that underwrites mnemic reference and the kind that underwrites remembering. However, this goes against the traditional causalist project of answering (Q1) and (Q2) (§2.1) simultaneously. Moreover, it involves agreeing that they do not have a readymade solution to the reference question, as is often assumed. The burden is then on the causalist to provide a suitable answer to the reference question.

[37] We hope to address these issues at greater length on a future occasion.

a priori, a dramatic decrease in the number of *genuine* confabulations.[38] In response to the second reply, the causalist told us that "confabulations lack a connection to any event in the confabulator's past" (Robins, 2020, p. 130). If this is merely an incidental feature, then what is at the heart of confabulation? The causalist must say more.

7 Conclusion

Traditional causal theories of remembering are motivated by the desire to explain both what remembering *is* and what *determines the referent* of each given occurrence of remembering. Even where they grant that memory processes are inherently constructive, they insist that there must, for genuine remembering to occur, always be a special ingredient (i.e., a memory trace) the causal aetiology of which *suffices to settle* the question of reference in episodic memory (§2.1). Yet there are pressing worries about the empirical credentials of this apparently simple picture (§12.2).

Postcausal theories are a product of caution about the necessity of such appropriate causation. It is natural for these theories to explain what remembering *is* and what makes it *refer* in different ways. They may tell us that to be remembering is to have a representation that was produced in certain distinctive way (Michaelian, 2016b), or that it is to be in a mental state that tends to be appropriately caused (Fernández, 2019). Once we have this explanation of what it is to be remembering, we can then focus on downstream questions concerning the contingent, semantic or normative features of remembering (i.e., reference or accuracy). While we agree that there is no easy solution to these questions (§4), we have articulated a framework within which postcausal theorists can begin to address the reference question. On this picture, which event(s) one is remembering is a matter of which event(s) one's representation reliably characterises, given the specific process which led to its production.

While it remains a blueprint in its finer details (much like the fruitful approaches that characterise knowledge in terms of reliability, or the truth-conditions of counterfactuals in terms of closeness), our reliabilist approach has important virtues. First, it promises to be applicable to cases of episodic future thinking (§5.1), in accordance with the continuist idea that remembering and various forms of constructive imagining are underpinned by mechanisms of fundamentally the same kind (e.g., Addis (2020)). Second, by prising apart our account of what determines reference in remembering (and in some forms of imagining) from our account of what remembering *is*, we reveal the possibility that these two phenomena can come apart, in particular in cases of referential confabulation (§6). According to our account, referential confabulations are products of improperly functioning episodic memory systems that nonetheless, on that occasion, produce a representation which fulfils the (precisified) referentiality condition.

Causal theories of remembering will struggle to possess these two virtues. If appropriate causation is needed for reference in remembering, reference in episodic future

[38] Schnider & Ptak (1999), who speculate that confabulation can result from "an inability to suppress activated memory traces" that are contextually irrelevant (680), note that, "[w]hether the stories seem simple or fantastic, they can virtually always be traced back to fragments of actual experiences" (677).

thought cannot be explained in the same way. Moreover, if it is sufficient for remembering that one now represent a particular event, that one experienced the event when it occurred, and that one's current representation has its reference fixed by being appropriately caused by a memory trace, then there cannot be cases of confabulation which possess these same features despite failing to be instances of genuine remembering. At the root of these worries for causal theories is that they use the notion of appropriate causation to pull off too many simultaneous feats: it is said to be the key mechanism that underpins remembering; to fully account for mnemic reference-fixing; to explain the reliability of remembering; to distinguish remembering from relearning, and; to distinguish remembering from mnemic confabulation.

Far from the reference question presenting an insurmountable challenge for postcausal theories of remembering, we have argued that these theories—and simulationism, in particular—have the beginnings of a plausible account, and that this account illuminates a neglected class of mnemic phenomena (namely, referential confabulations) which present a serious challenge to causal theories of remembering.

Acknowledgements This material benefitted from presentations at the *Reference in Remembering* workshop, *Issues in Philosophy of Memory 3*, the Federal University of Rio Grande do Sul, and the 7th meeting of the *Brazilian Society for Analytic Philosophy*. Thanks to Sara Aronowitz, Michael Barkasi, José Carlos Camillo, Imogen Dickie, Megan Entwistle, Peter Langland-Hassan, Denis Perrin, André Sant'Anna, and Markus Werning for helpful questions and discussions. We are also particularly grateful to Ali Boyle and an anonymous reviewer for this journal for helpful written comments.

Funding James Openshaw's work on this project received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101032391. It was also funded by Grant Nos. 397530566 and 419038924 from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the DFG research group Constructing Scenarios of the Past (FOR 2812). Kourken Michaelian's work on this project was supported by the French National Research Agency ("Investissements d'avenir" program ANR-15-IDEX-02 and grant ANR-23-CE54-0004-01), by CAPES-COFECUB (grant Sh 967/20), and by ORCHID (grant 112-2927-I-A49A-501).

Declarations

Conflict of interest The authors hereby state that there are no conflicts of interest to declare.

References

Addis, D. R. (2020). Mental time travel? A neurocognitive model of event simulation. *Review of Philosophy & Psychology, 11*, 233–259.

Andonovski, N. (2022). Causation in memory: Necessity, reliability and probability. *Acta Scientiarum, 43*(3), e61493.

Aronowitz, S. (2019). Memory is a modelling system. *Mind & Language, 34*(4), 483–502.

Baddeley, A., & Wilson, B. (1986). Amnesia, autobiographical memory and confabulation. In D. C. Rubin (Ed.), *Autobiographical memory.* Cambridge University Press.

Barkasi, M. (Forthcoming). Consumer-side reference through promiscuous memory traces, *Synthese*.

Berlyne, N. (1972). Confabulation. *British Journal of Psychiatry, 120*, 31–39.

Bernecker, S. (2010). *Memory: A Philosophical Study*. Oxford University Press.

Bernecker, S. (2015). Visual memory and the bounds of authenticity. In A. Coliva, V. Munz, & D. Moyal-Sharrock (Eds.), *Mind, language and action: Proceedings of the 36th international Wittgenstein symposium.* De Gruyter.

Bernecker, S. (2017). A causal theory of mnemonic confabulation. *Frontiers in Psychology, 8*, 1207.

Berntsen, D. (2021). Involuntary autobiographical memories and their relation to other forms of spontaneous thoughts. *Philosophical Transactions: Biological Sciences, 376*(1817), 1–9.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex, 19*, 2767–2796.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences, 15*(11), 527–536.

Bortolotti, L., & Cox, R. E. (2009). Faultless' ignorance: Strengths and limitations of epistemic definitions of confabulation. *Consciousness and Cognition, 18*, 952–965.

Burgess, P. W., & Shallice, T. (1996). Confabulation and the control of recollection. *Memory, 4*(4), 359–411.

Campbell, J. (2002). *Reference and consciousness*. Clarendon Press.

Campbell, S. (2006). Our faithfulness to the past: Reconstructing memory value. *Philosophical Psychology, 19*(3), 361–380.

Dalla Barba, G., Cipolotti, L., & Denes, G. (1990). Autobiographical memory loss and confabulation in Korsakoff's syndrome: A case report. *Cortex, 26*, 525–534.

Dalla Barba, G. (2002). *Memory, consciousness and temporality*. Kluwer Academic Publishers.

Dalla Barba, G. (2009). Temporal consciousness and confabulation: Escape from unconscious explanatory idols. In W. Hirstein (Ed.), *Confabulation: Views from neuroscience, psychiatry, psychology and philosophy.* Oxford University Press.

De Brigard, F. (2014). Is memory for remembering? Recollection as a form of episodic hypothetical thinking. *Synthese, 191*(2), 155–185.

Deutsch, M. (2021). Is there a 'qua problem' for a purely causal account of reference grounding? *Erkenntnis, 1*, 1–18.

Devitt, M. (1981). *Designation*. Columbia University Press.

Dickie, I. (2015). *Fixing reference*. Oxford University Press.

Dickie, I. (2017). Reply to Hofweber and Ninan. *Philosophy and Phenomenological Research, 95*(3), 745–760.

Evans, G. (1973). The causal theory of names. *Aristotelian Society Supplementary, 47*(1), 187–208.

Evans, G. (1982). *The varieties of reference*. Clarendon Press.

Fayyaz, Z., Altamimi, A., Zöllner, C., Klein, N., Wolf, O. T., Cheng, S., & Wiskott, L. (2022). 'A model of semantic completion in generative episodic memory. *Neural Computation, 34*(9), 1841–1870.

Fernández, J. (2019). *Memory: A self-referential account*. Oxford University Press.

Ghosh, V. E., & Gilboa, A. (2014). 'What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia, 53*, 104–114.

Goldman, A. (1976). Discrimination and perceptual knowledge. *The Journal of Philosophy, 73*(20), 771–791.

Gonsalves, B., Reber, P. J., Gitelman, D. R., Parrish, T. B., Mesulam, M. M., & Paller, K. A. (2004). Neural evidence that vivid imagining can lead to false remembering. *Psychological Science, 15*(10), 655–660.

Grundmann, T. (2018). Saving safety from counterexamples. *Synthese, 197*(12), 5161–5185.

Hawthorne, J. (2003). *Knowledge and lotteries*. Oxford University Press.

Hawthorne, J., & Manley, D. (2012). *The reference book*. Oxford University Press.

Henkel, L. A., & Franklin, N. (2000). Cross-modal source monitoring confusions between perceived and imagined events. *Journal of Experimental Psychology, 26*(2), 321–335.

Hirstein, W. (2005). *Brain Fiction*. MIT Press.

Hoerl, C. (2018). Remembering past experiences: Episodic memory, semantic memory, and the epistemic asymmetry. In K. Michaelian, D. Debus, & D. Perrin (Eds.), *New directions of research in the philosophy of memory.* Routledge.

Hopkins, R. (2018). Imagining the past: On the nature of episodic memory. In F. Macpherson & F. Dorsch (Eds.), *Memory and imagination.* Oxford University Press.

Hutto, D. D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. MIT Press.

Jeshion, R. (2010). Singular thought: Acquaintance, semantic instrumentalism, and cognitivism. In R. Jeshion (Ed.), *New essays on singular thought.* Oxford University Press.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*(1), 3–28.

Kaplan, D. (1969). Quantifying. In D. Davidson & J. Hintikka (Eds.), *Words and objections.* Reidel.

Kaplan, D. (1989). Afterthoughts. In J. Almog, J. Perry, & H. Wettstein (Eds.), *Themes from Kaplan.* Oxford University Press.

Kripke, S. (1980). *Naming and necessity*. Harvard University Press.

Langland-Hassan, P. (2022). Propping up the causal theory. *Synthese, 200*(2), 1–27.

Lewis, D. K. (1979). Counterfactual dependence and time's arrow. *Noûs, 13*(4), 455–476.

Lewis, D. K. (1986). *'A subjectivist's guide to objective chance', in his Philosophical Papers* (Vol. 2). Oxford University Press.

Martin, C. B., & Deutscher, M. (1966). Remembering. *The Philosophical Review, 75*(2), 161–196.

McCarroll, C. J. (2018). *Remembering from the outside: Personal memory and the perspectival mind*. Oxford University Press.

McCarroll, C. J., Michaelian, K., & Nanay, B. (2022). Explanatory contextualism about episodic memory: Towards a diagnosis of the causalist-simulationist debate. *Erkenntnis*. https://doi.org/10.1007/s10670-022-00629-4

Michaelian, K. (2016a). Confabulating, misremembering, relearning: The simulation theory of memory and unsuccessful remembering. *Frontiers in Psychology, 7*, 1857.

Michaelian, K. (2016b). *Mental time travel: Episodic memory and our knowledge of the personal past*. MIT Press.

Michaelian, K. (2020). Confabulating as unreliable imagining: In defence of the simulationist account of unsuccessful remembering. *Topoi, 39*(1), 133–148.

Michaelian, K. (2021). Imagining the past reliably and unreliably: Towards a virtue theory of memory. *Synthese, 199*(3–4), 7477–7507.

Michaelian, K. (2023). Towards a virtue-theoretic account of confabulation. In A. Sant'Anna, C. McCarroll, & K. Michaelian (Eds.), *Current controversies in philosophy of memory.* Routledge.

Michaelian, K. (2022). Radicalizing simulationism: Remembering as imagining the (nonpersonal) past. *Philosophical Psychology*. https://doi.org/10.1080/09515089.2022.2082934

Michaelian, K., & Robins, S. K. (2018). Beyond the causal theory? Fifty years after Martin and Deutscher. In K. Michaelian, D. Debus, & D. Perrin (Eds.), *New directions of research in the philosophy of memory.* Routledge.

Michaelian, K., & Sant'Anna, A. (2021). Memory without content? Radical enactivism and (post)causal theories of memory. *Synthese, 198*(Suppl 1), S307–S335.

Michaelian, K., & Sant'Anna, A. (2022). From authenticism to alethism: Against McCarroll on observer memory. *Phenomenology and the Cognitive Sciences, 21*, 835–856.

Moscovitch, M. (1989). Confabulation and the frontal systems: Strategic versus associative retrieval in neuropsychological theories of memory. In H. L. Roediger III. & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving.* Lawrence Erlbaum Associates Inc.

Moscovitch, M. (1995). Confabulation. In D. L. Schacter (Ed.), *Memory distortion: How minds, brains, and societies reconstruct the past.* Harvard University Press.

Neisser, U. (1997). The ecological study of memory. *Philosophical Transactions: Biological Sciences, 352*(1362), 1697–1701.

Ninan, D. (2017). Aboutness and justification. *Philosophy and Phenomenological Research, 95*(3), 731–737.

Openshaw, J. (2021). Thinking about many. *Synthese, 199*, 2863–2882.

Openshaw, J. (2022). Remembering objects. *Philosophers' Imprint, 22*(11), 1–20.

Openshaw, J. (2023). (In defence of) preservationism and the previous awareness condition: What is a theory of remembering, anyway? *Philosophical Perspectives, 37*(1), 290–307.

Openshaw, J. (2022). Does singular thought have an epistemic essence? *Inquiry*. https://doi.org/10.1080/0020174X.2022.2155871

Pepp, J. (2020). Is Dickie's account of aboutness-fixing explanatory? *Theoria, 86*, 801–820.

Perrin, D. (2016). Asymmetries in subjective time. In K. Michaelian, S. B. Klein, & K. K. Szpunar (Eds.), *Seeing the future: Theoretical perspectives on future-oriented mental time travel.* Oxford University Press.

Perrin, D. (2021). Embodied episodic memory: A new case for causalism? *Intellectica, 74*, 229–252.

Prasad, D., & Bainbridge, W. A. (2023). The visual Mandela effect as evidence for shared and specific false memories across people. *Psychological Science*. https://doi.org/10.1177/09567976221108944

Recanati, F. (2007). *Perspectival thought: A plea for (moderate) relativism*. Oxford University Press.

Recanati, F. (2012). *Mental files*. Oxford University Press.

Renoult, L., Irish, M., Moscovitch, M., & Rugg, M. D. (2019). From knowing to remembering: The semantic-episodic distinction. *Trends in Cognitive Science, 23*(12), 1041–1057.

Robins, S. K. (2016). Representing the past: Memory traces and the causal theory of memory. *Philosophical Studies, 173*(11), 2993–3013.

Robins, S. K. (2019). Confabulation and constructive memory. *Synthese, 196*, 2135–2151.

Robins, S. K. (2020). Mnemonic confabulation. *Topoi, 39*, 121–132.

Rubin, D. C. (2022). A conceptual space for episodic and semantic memory. *Memory & Cognition, 50*, 464–477.

Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions: Biological Sciences, 362*(1481), 773–786.

Schnider, A. (2018). *The confabulating mind: How the brain creates reality*. Oxford University Press.

Schnider, A., & Ptak, R. (1999). Spontaneous confabulators fail to suppress currently irrelevant memory traces. *Nature Neuroscience, 2*(7), 677–681.

Sterelny, K. (1990). *The representational theory of mind: An introduction*. Basil Blackwell.

Strickland, B., & Keil, F. (2011). Event completion: Event based inferences distort memory in a matter of seconds. *Cognition, 121*(3), 409–415.

Strikwerda-Brown, C., Shaw, S. R., Hodges, J. R., Piguet, O., & Irish, M. (2022). Examining the episodic-semantic interaction during future thinking: A reanalysis of external details. *Memory & Cognition, 50*, 617–629.

Soteriou, M. (2018). The past made present: Mental time travel in episodic recollection. In K. Michaelian, D. Debus, & D. Perrin (Eds.), *New directions of research in the philosophy of memory.* Routledge.

Talland, G. A. (1965). *Deranged memory. A psychonomic study of the Amnesic Syndrome*. Academic Press.

Tolly, J. (2021). Knowledge, evidence, and multiple process types. *Synthese, 198*, 5625–5652.

Werning, M. (2020). Predicting the past from minimal traces: Episodic memory and its distinction from imagination and preservation. *Review of Philosophy and Psychology, 11*(2), 301–333.

Werning, M., & Liefke, K. (2023). Remembering dreams: Parasitic reference in memories of non-veridical experiences. In D. Gregory & K. Michaelian (Eds.), *Dreaming and memory: Philosophical issues.* Springer.

Williams, J. R. G. (2008). Gavagai again. *Synthese, 164*(2), 235–259.

Williamson, T. (2009). Probability and danger. *The Amherst Lecture in Philosophy, 4*, 1–35.

Zöllner, C., Klein, N., Cheng, S., Schubotz, R. I., Axmacher, N., & Wolf, O. T. (2023). Where was the toaster? A systematic investigation of semantic construction in a new virtual episodic memory paradigm. *Quarterly Journal of Experimental Psychology, 76*(7), 1497–1514.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.